Modeling Correlated Binary Data in Clinical Trials

Mohamed Al-Osh, Ph.D.
Division of Biometrics II, CDER, FDA*

A Draft for a Presentation at the
ASA Annual Meeting in Baltimore, August 1999

* The views expressed herein are those of the author and not the FDA.

An Outline

I. Introduction/ Motivation

II. Some properties of binary variates

III. An approach for generating and modeling correlated binary data

IV. Modeling multiple correlated binary measurements,

-An application to diagnostic testing

-Improving the fit by introducing further dependence among the multiple tests

Introduction/ Motivation:

II. Modeling Correlated Binary Data:

II.A : Correlation due to sharing some common element (X).

Examples: Measurements on:

-pair of eyes or ears (same person), or

- on siblings (same parents), or

- on tooth's decay (same location: mouth).

II.B. Possible correlation due to similarity in the mechanism that generated the data, as in diagnostic tests.

An approach for modeling and generating multiple correlated binary data is desired to:

- investigate small sample properties of estimation methods such as the GEE method.
- model and analyze such data.

II. Some Properties of the Binary Variates:

Property II.1

Let X and U be two indep. r.v. s.t. $X \sim$ Ber (α) and

$U \sim$ Ber (β), then:

$$Y = UX \tag{2.1}$$

Then: $Y \sim$ Ber ($\alpha\ \beta$), and $1\text{-}Y \sim$ Ber ($1\text{-}\alpha\ \beta$)

Property II.2:

Let U and V be two indep. r.v. s.t. $U \sim$ Ber(β)

and$V \sim$ Ber($1\text{-}\theta$) and X be as in II.1, and define:

$$Y = U X + V (1\text{-}X) \tag{2.2}$$

Then: $Y \sim \text{Ber}[\alpha\, \beta + (1\text{-}\alpha)\, (1\text{-}\theta)]$.

That is, the mixture of two binary variates is again a binary variate.

Interpretation:

Let X be the true unobserved disease status of a patient and let Y be the results of an error-prone test, then by (2.2) we have:

$$P(\ Y = 1) = \beta\, \alpha + (1-\theta)(1-\alpha)$$

and $$P(\ Y = 0) = (1-\beta)\, \alpha + \theta(1-\alpha)$$

In evaluating the accuracy of a diagnostic test two types of errors are usually encountered:

$P(\ Y = 0/\ X =1) = \beta$ FNR (=1- sensitivity)

$P(\ Y = 1/\ X =0) = \theta$ FPR (=1- specificity)

A similar interpretation holds for signal transmission.

We will re-visit the above interpretation for diagnostic testing in the application (Section IV).

Property II.3:

Let $\{ U_i \}_{i=1}^{m}$ be a sequence of indep. binary r.v.'s

with parameters β_i , i=1,2, ...k, then:

$$Y = \prod_{i=1}^{m} U_i$$

is again a binary r.v. with parameter

$(\prod_{i=1}^{m} \beta_i)$, denoted as $Y \sim$ Ber $(\prod_{i=1}^{m} \beta_{ij})$.

Properties II.1 and II.2 can be used for generating pairs of correlated binary data, and Property II.3 can be used for generating a vector of arbitrary dimensions of correlated binary variates.

For modeling, X plays the role of the common element, which induces the correlation between the binary data.

III. Generating Pairs of Correlated Binary Variates:

Use Property II.1, to define $Y_{i\,j}$ as:

$Y_{i\,j} = U_{ij} X_i$ for $i=1,2, ... k$; and $j=1,2$ (3.1)

where X_i $(i=1,2, ... k)$ is a set of indep. Ber. (α_i) variates and U_{ij} $(i=1,2, ... k, j=1,2)$ is a set of indep.

Ber. (β_{ij}) variates which are independent also of the X_i's. Then by Property (II.1), we have:

$$\mathscr{E}(Y_{ij}) \equiv p_{ij} = \alpha_i \beta_{ij} \qquad (3.1)$$

$$\rho_{i12} = (1-\alpha_i)\, p_{i1}\, p_{i2} / \alpha_i\, \sigma_{i1}\, \sigma_{i2} \qquad (3.2)$$

ρ_{i12} satisfies the following bounds :

$$0 \leq \rho_{i12} \leq \min \{ (p_{i1} q_{i2} / q_{i1} p_{i2})^{½}, (q_{i1} p_{i2} / p_{i1} q_{i2})^{½} \} \qquad (3.3)$$

For a given set of (p_{ij}, j=1,2) and (ρ_{i12}'s), one can solve (3.2) and (3.3) for the set of parameters α_i, β_{i1}, and β_{i2} in order to generate correlated variates with the required values for p_{ij} and ρ_{i12}'s. Specifically we have:

$$\alpha_i = [\ p_{i1} p_{i2} / (\rho_{i12} \sigma_{i1} \sigma_{i2} + p_{i1} p_{i2})]$$

and $\beta_{ij} = p_{ij} / \alpha_i$ for j = 1, 2

III. 2 Generating Pairs of Non-positively Correlated Binary Variates:

Use of properties II.1 and II.2 and define:

$$Y_{i1} = U_{i1} X_i$$
$$Y_{i2} = 1\text{-}U_{i2} X_i \qquad \text{for } i=1,2, \dots k \qquad (3.4)$$

Here ρ_{i12} is bounded by:

$$\max \{- (p_{i1}\, p_{i2}/q_{i1}\, q_{i2})^{½}, - (q_{i1}\, q_{i2}/p_{i1}\, p_{i2})^{½}\}$$
$$\leq \rho_{i12} \leq 0 \qquad (3.5)$$

III. 3 Generating Pairs of Correlated Binary Variates with Full Range Correlation:

Use Property II.3, and define:

$$Y_{ij} = V_{ij} U_{ij} X_i + (1\text{-}V_{ij}) (1\text{-}U_{ij} X_i)$$
$$\text{for } i=1,2, \dots k \quad \text{and } j=1,2 \qquad (3.6)$$

where X_i and $U_{i\,j}$ (i=1,2, ... k; j=1,2) as defined in III.1 and V_{ij} is a sequence of indep. Bern.(θ_{ij}) rv, which are indep. of $X_{\,i}$ and $U_{i\,j}$ (i=1,2, ... k; j=1,2).

The representation in (3.6) reduces to that of (3.1) for $\theta_{\,i1}$ =1 and $\theta_{\,i2}$ =1 and it reduces to that of (3.4) for $\theta_{\,i1}$ =1 and $\theta_{\,i2}$ =0 .

$$\rho_{i12} = \alpha_{\,i}\ (1-\alpha_{\,i}\,)\ \beta_{i\,1}\beta_{\,i\,2}\ (2\theta_{i1}- 1)\ (2\theta_{\,i2} - 1)\ /\sigma_{i\,1}\ \sigma_{i2} \qquad (3.7)$$

ρ_{i12} is non-negative when each of θ_{i1} and θ_{i2} > (<) 0.5 ; and it is negative when $\theta_{i1} > 0.5$ and $\theta_{i2} < 0.5$ or vice versa. ρ_{i12} satisfies:

$$\max \{- (p_{\,I\,1}\, p_{\,i2}/q_{\,I\,1}\, q_{\,i2}\,)^{\,½}, - (q_{\,I\,1}\, q_{\,i2}/p_{\,I\,1}\, p_{\,i2}\,)^{\,½}\} \le \rho_{\,i12}$$

$$\le \min \{ (p_{\,I\,1}\, q_{\,i2}/q_{\,I\,1}\, p_{\,i2}\,)^{\,½}, (q_{\,I\,1}\, p_{\,i2}/p_{\,I\,1}\, q_{\,i2}\,)^{\,½}\} \qquad (3.8)$$

The range of ρ_{i12} in (3.8) is the max. (Prentice, 1988, and Emrich and Piedmonte, 1991).

For a given set of p_{ij}'s and ρ_{i12}'s, one can use (3.6) to generate k pairs of correlated binary variates.

IV. Application:

IV.I. HIV data (Qu et el.,1996, Yang & Becker, 1997), results of 4 diag. tests applied to 428 HIV patients.

Table 1: Freq. & Res. of Fitted LCM to 4 Tests Class. of 428 HIV Patients [1]

Response Pattern				Frequency	Residuals	
Y1 Z1	Y2 Z2	Y3 Z3	Y4 Z4		L C M	LCM + λ 23
1	1	1	1	128	7.547	0.075
1	1	1	0	0	-0.024	-0.019
1	1	0	1	4	-7.531	-0.046
1	1	0	0	1	0.809	0.820
1	0	1	1	83	-7.486	0.211
1	0	1	0	0	-0.066	-0.015
1	0	0	1	17	7.902	-0.193
1	0	0	0	4	-1.129	-0.811
0	1	1	1	0	-0.017	-0.012
0	1	1	0	0	-0.058	-0.001
0	1	0	1	0	-0.542	-0.536
0	1	0	0	6	-0.182	-0.276
0	0	1	1	0	-0.148	-0.011
0	0	1	0	0	-1.589	-0.025
0	0	0	1	15	0.307	0.537
0	0	0	0	170	2.209	0.302

[1] Data Source: Qu et al (Biometrics, 1996, 798-808)

The purpose of the analysis is to estimate the accuracy of each of the diagnostic tests in the absence of the ‘gold standard’, after accounting for the dependence when it is present.

Analysis Steps:

- Consider the representation in (2.2) for each test, and assume given X=x, the tests are independent. This is the classical setting for the Latent Class Model (LCM).

$$L(\beta, \theta, \alpha / z) \propto$$

$$\prod_{z} \left[\prod_{i=1}^{k} \beta_i^{\,1-z_i} (1-\beta_i)^{z_i} + (1-\alpha) \prod_{i=1}^{k} \theta_i^{\,z_i} (1-\theta_i)^{1-z_i} \right]^{n(z)}$$

A nonlinear optimization algorithm can be used to derive the MLE of the parameters of the model. The results of this fit are given in the following table.

Par.	Estimate	(Asym.S.E.)	Asy. 95 % C.I.
α	0.540	(0.024)	(0.483 , 0.597)
$\beta1$	0.000	(0.001)	(-.002 , 0.003)
$\beta2$	0.429	(0.033)	(0.352, 0.506)
β 3	0.087	(0.019)	(0.043 , 0.132)
$\beta4$	0.000	(0.001)	(-.003 , 0.003)
$\theta1$	0.030	(0.013)	(-.000 , 0.060)
$\theta2$	0.036	(0.013)	(0.004 , 0.067)
$\theta3$	0.009	(0.007)	(-.007 , 0.026)
$\theta4$	0.081	(0.020)	(0.034 , 0.127)

Goodness of Fit:

Source	**SS**	**DF**	**Weighted MS**
Residuals	19.031	7	2.719

Examination of the residuals from the fitted LCM

(Table 1) shows dependency between tests 2 and 3 results, when the diagnoses of the tests are positive. To accommodate this dependence we extend the LCM by including a dependence parameter (r_β) in the model (Vacek, 1985 and Torrance-Rynard and Walter, 1996).

When the underlying true diagnosis is positive, the dependence parameter between tests 2 and 3 (r_β) is bounded by:

$r_\beta \leq \beta_2 (1-\beta_3)\beta_1 \beta_4$ and $r_\beta \leq (1-\beta_2) \beta_3 \beta_1 \beta_4$

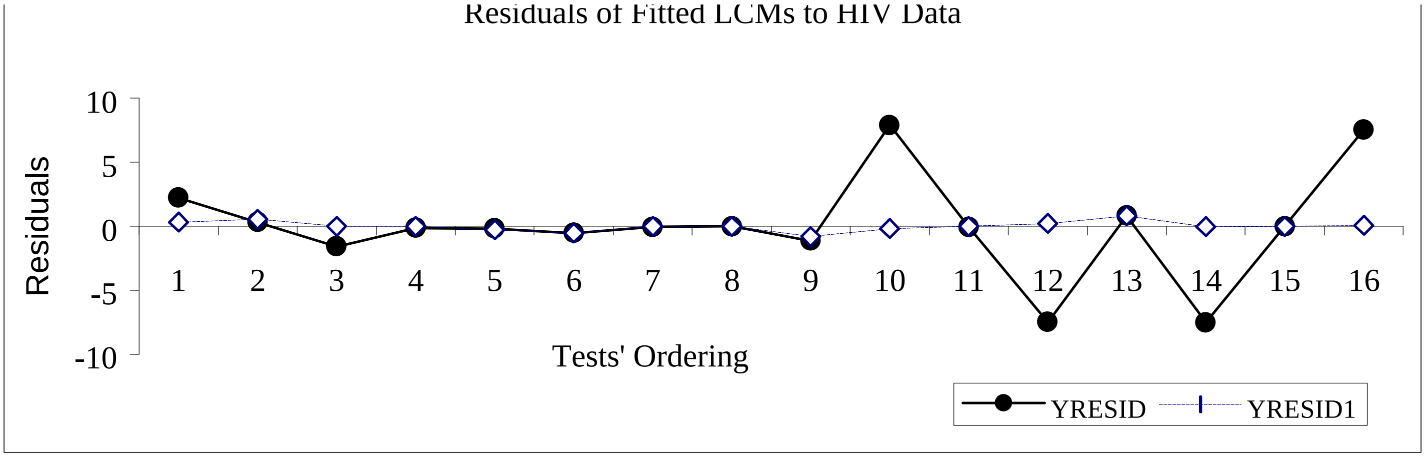
A similar relation holds when the true diagnosis is negative r_θ .

Results of including both dependencies r_β and r_θ show that the contribution of r_θ to improving the fit, in the presence of r_β , is minimal. Thus, we include only r_β. The results of fitting this model are given in the following table.

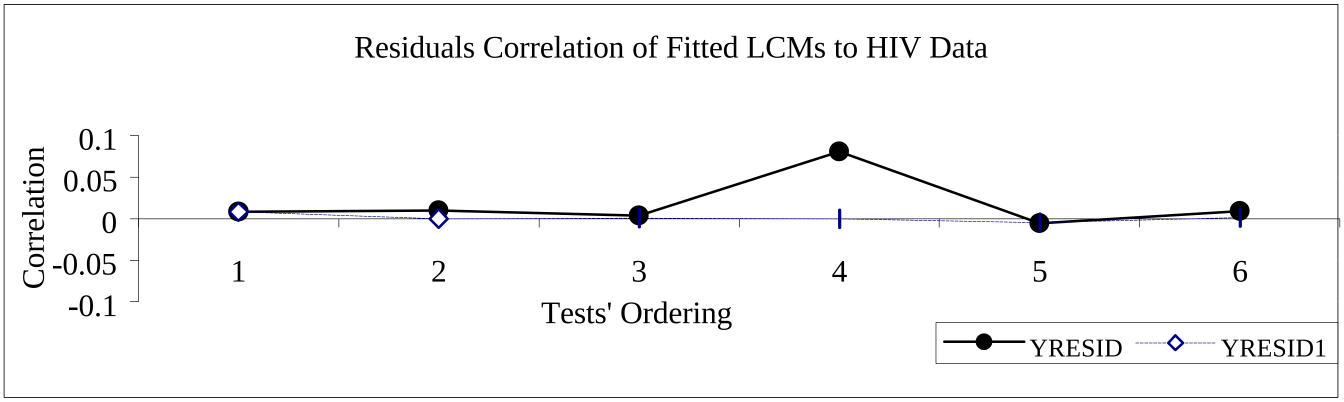
Par.	Estimate	(Asym.S.E.)	Asy. 95 % C.I.
C	0.660	(0.149)	(0.295 , 1.025)
α	0.541	(0.024)	(0.482 , 0.600)
$\beta 1$	0.000	(0.001)	(-.002 , 0.002)
$\beta 2$	0.430	(0.033)	(0.350, 0.510)
β 3	0.090	(0.019)	(0.043 , 0.136)
$\beta 4$	0.000	(0.001)	(-.002 , 0.002)
$\theta 1$	0.028	(0.012)	(-.002 , 0.057)
$\theta 2$	0.036	(0.013)	(0.003, 0.068)
$\theta 3$	0.000	(0.001)	(-.002, 0.002)
$\theta 4$	0.079	(0.020)	(0.031 , 0.126)

Goodness of Fit:

Source	**Weighted SS**	**DF**	**Weighted MS**



Residuals 4.533 6 0.756



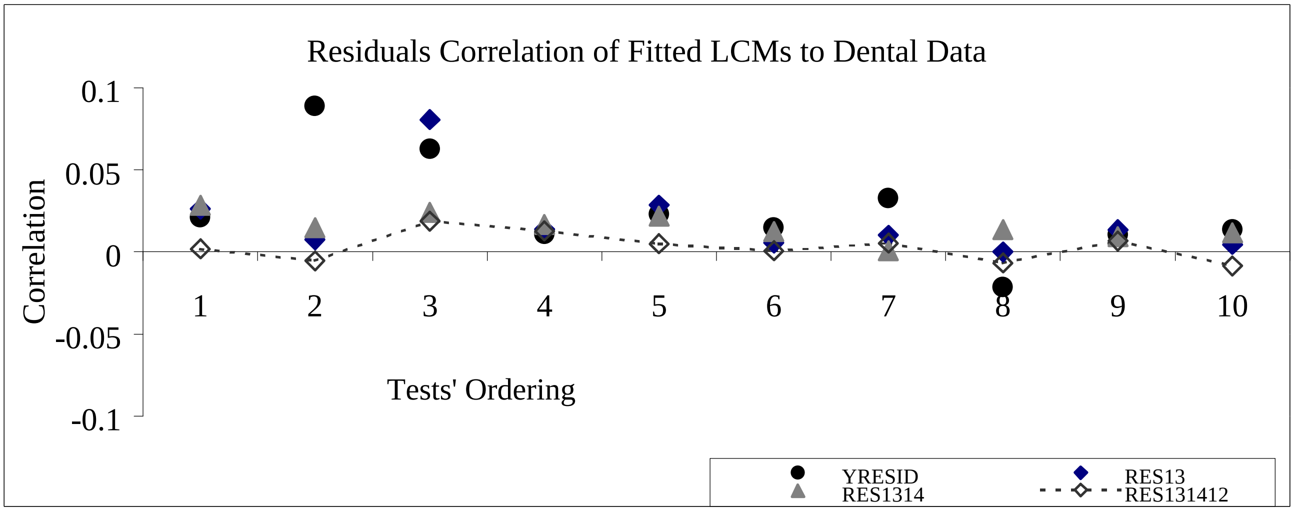
IV.II. Dentistry Data:(Espeland and Hanelman,1989)

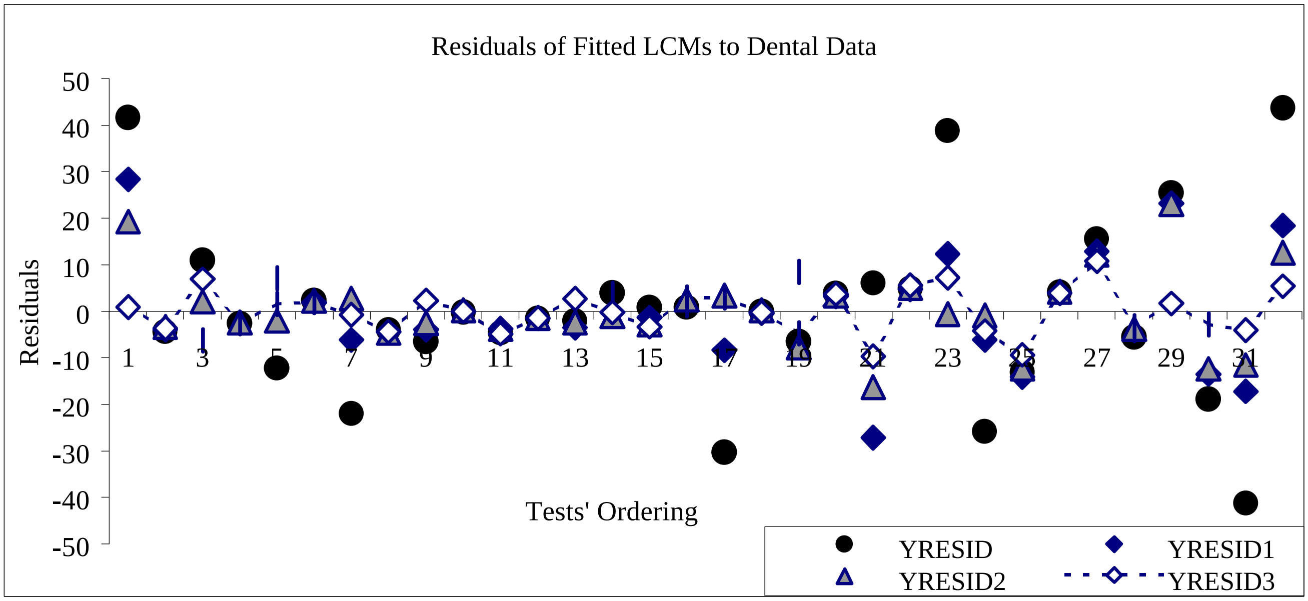
Table 2: Freq. and Res. of Fitted LCM to 5 dentists Class.of 3869 denX-Ray [1]

Response Pattern					Frequency	Residuals			
Y1	Y2	Y3	Y4	Y5		L C M	LCM +λ13	LCM + λ13 + λ14	LCM + λ13 + λ14 + λ12
1	1	1	1	1	100	41.614	28.312	19.219	0.913
1	1	1	1	0	1	-4.392	-3.939	-3.439	-3.703
1	1	1	0	1	72	10.936	-6.302	2.059	6.851
1	1	1	0	0	3	-2.668	-2.611	-2.408	-2.601
1	1	0	1	1	27	-12.261	7.006	-2.040	1.594
1	1	0	1	0	6	2.305	1.882	2.273	2.011
1	1	0	0	1	20	-22.002	-6.191	2.740	-0.807
1	1	0	0	0	2	-4.023	-4.483	-4.642	-4.421
1	0	1	1	1	17	-6.519	-4.015	-2.554	2.312
1	0	1	1	0	2	-0.180	-0.071	0.205	-0.081
1	0	1	0	1	20	-4.708	-3.870	-3.819	-4.830
1	0	1	0	0	1	-1.542	-1.569	-1.446	-1.478
1	0	0	1	1	14	-2.082	-3.515	-2.409	2.693
1	0	0	1	0	6	3.900	3.760	-1.115	-0.335
1	0	0	0	1	26	0.830	-1.380	-2.880	-3.348
1	0	0	0	0	22	0.791	1.276	2.499	2.880
0	1	1	1	1	56	-30.407	-8.354	3.327	2.947
0	1	1	1	0	8	-0.068	-0.487	0.178	-0.271
0	1	1	0	1	85	-6.570	8.448	-7.760	-4.727
0	1	1	0	0	15	3.792	3.154	3.355	3.689
0	1	0	1	1	67	5.988	-27.176	-16.316	-9.725
0	1	0	1	0	17	4.854	4.696	4.977	5.489
0	1	0	0	1	191	38.795	12.274	-0.670	7.201
0	1	0	0	0	188	-25.893	-6.230	-0.854	-4.276
0	0	1	1	1	22	-13.147	-14.101	-12.446	-9.396
0	0	1	1	0	8	3.965	3.814	4.163	3.816
0	0	1	0	1	63	15.523	12.850	12.089	10.765
0	0	1	0	0	23	-5.628	-4.114	-3.848	-3.236
0	0	0	1	1	75	25.362	23.234	23.011	1.618
0	0	0	1	0	43	-18.932	-13.491	-12.455	-2.820
0	0	0	0	1	789	-41.353	-17.221	-11.551	-4.061
0	0	0	0	0	1880	43.720	18.413	12.557	5.337

Goodness of Fit for a sequence of Models:

Source	Weighted SS	DF	Weighted MS
Res. (LCM)	131.997	21	6.286
Res. (LCM_{13})	74.104	20	3.705
Res. (LCM_{1314})	49.298	19	2.595
Res. (LCM_{131412})	27.712	18	1.540




Residuals of Fitted LCMs to Dental Data
Residuals
50
40
30
20
10
0
-10
-20
-30
-40
-50
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
Tests' Ordering
YRESID
YRESID1
YRESID2
YRESID3